Korpuslinguistik mit Online-Ressourcen
Workshop des LIPP-Doktorandenkolloquiums (LMU München)

Stefanie Dipper, Stefan Evert, Heike Zinsmeister

28. Januar 2011

http://wordspace.collocations.de/doku.php/corpus_tutorial:lipp2011

Überblick

1. Europarl — einfache Anfragen ("Simple Query"-Modus und "Word List"-Modus)
2. Europarl — CQPweb ("CQP Mode")
3. DWDS-Suche: eine diachrone Perspektive
4. Chunks und morphologische Merkmale

1 Europarl – einfache Anfragen ("Simple Query"-Modus und "Word List"-Modus)

In der ersten Übung untersuchen Sie das Europarl-Korpus (Debatten des europäischen Parlaments aus den Jahren 1996–2006) mit einem speziell für dieses Korpus entwickelten Webinterface. Sie können dabei die einfache Anfragesprache CEQL ("Common Elementary Query Language") verwenden, die auch von einigen anderen Web-Interfaces unterstützt wird.

Geben Sie zuächst in Ihren Browser die Adresse des Europarl-Webinterfaces in Osnabrück ein. Auf der Startseite finden Sie in der Menüzeile unter dem Korpusnamen drei Links auf weiterführende Seiten. Wählen Sie hier die Option `simple query`. Auf der Anfrageseite zur *Simple Query* finden Sie einen Link auf eine Hilfeseite (`Help page` im Hauptmenü), auf der die Syntax der einfachen Anfragesprache erläutert wird.

- **URL:** http://cogsci.uni-osnabrueck.de/~korpora/ws/CQPdemo/Europarl/
  ▶ `simple query`
- **Hilfe zur einfachen Anfragesprache:**
  http://cogsci.uni-osnabrueck.de/~korpora/ws/CQPdemo/Europarl/help-simple.html

1.1 Suche nach einem Einzelwort

Um die Anfragesprache und den Inhalt des Korpus kennenzulernen, suchen Sie zunächst nach einem einfachen Wort. Wählen Sie bei der Spracheinstellung `German` und als Sprachen für das Ausgabe-Display `EN, DE` (Englisch, Deutsch). Tippen Sie dann in das Anfragefenster `deutsch` und klicken Sie auf den Knopf `Run Query`. Welche Treffer erhalten Sie? Betrachten Sie sich den dritten Treffer genauer. Hier finden Sie den seltsamen Ausdruck *Eingeborenenvölker*. Lesen Sie den textuellen Kontext für diesen Treffer (Klick auf `context`). Wie wird *Eingeborenenvölker* auf Englisch übersetzt? Welcher Begriff wurde ursprünglich benutzt?

- **Einstellungen:** `lang=German; Display: EN, DE`; Rest wie voreingestellt
- **Anfrage:** `deutsch` (▶ `Run Query`) [R: 142]
- **Aufgabe:** Kontextanzeige für den 3. Treffer (▶ `context`)

Eigentlich ist es unerwartet, dass auch Treffer mit großgeschriebenem *Deutsch* gefunden werden. Um das zu verhindern, müssen Sie den Suchmodus auf wortgetreue Suche einstellen (`mode=literal`). Suchen Sie erneut und überprüfen Sie die Suchergebnisse.

- **Einstellung:** `mode=literal`
- **Erneute Anfrage:** `deutsch` (▶ `Run Query`) [R: 37]

1.2 Suche nach einer idiomatischen Wendung und ihrer Übersetzung

Der erste Treffer der letzten Suche weist eine interessante Wendung auf: *auf gut deutsch*. Wie wird diese Wendung übersetzt? Suchen Sie hierfür nach der gesamten Wendung (mit variabler Groß/Kleinschreibung). Wie viele Treffer erhalten Sie?

Lassen Sie sich zusätzlich zum Deutschen und Englischen die Treffer in Französisch (`FR`) und Niederländisch (`NL`) anzeigen. Die Originalsprache des Beitrags ist oft in der Kopfzeile nach dem Sprechernamen angegeben, d.h. die Treffer dokumentieren sowohl Übersetzungen aus dem Deutschen als auch Übersetzungen *in* das Deutsche.

- **Einstellung:** `lang=German; mode=case-folded; Display: EN, DE, FR, NL;`
- **Anfrage:** `auf gut deutsch` (▶ `Run Query`) [R: 8]
- **Aufgabe:** Finden Sie die Übersetzungen für *auf gut deutsch* bzw. Quellen, die ins Deutsche mit *auf gut deutsch* übersetzt wurden.

1.3 Suche nach einer flektierten Wortart: attributives Adjektiv

Wir wollen nun nach Verwendungen von *deutsch* als attributivem Adjektiv suchen. Bisher haben wir nur nach einer spezifischen Wortform gesucht, vgl. die erste Anfrage in Abschnitt 1.1 nach `deutsch`. Dieser Typ von Anfrage ist für flektierte Formen nicht hilfreich. Stattdessen verwenden Sie wieder die Einstellung `mode=literal` und erweitern das Suchmuster mit Platzhalteroperatoren (`*`,`+`) für die Flexionsendung des Adjektivs. Betrachten Sie jeweils die Ergebnisse. Lassen Sie sich dann die Treffer alphabetisch sortieren (Suche mit `sort order=ascending`). Was fällt Ihnen auf?

Bei der dritten Anfrageoption erhalten Sie 2890 Treffer. Lassen Sie sich hierfür auch die Häufigkeiten aller gefundenen Wortformen anzeigen (▶ `Frequencies`).

- **Einstellung:** `mode=literal`
- **Anfrage 1:** `deutsch*` (Endung optional) [R: 3046]
- **Aufgabe:** Sortieren Sie Wörter alphabetisch: `sort order=ascending`
- **Anfrage 2:** `deutsch+` (Fortsetzung obligatorisch) [R: 3009]
- **Anfrage 3:** `deutsche*` (Endung obligatorisch) [R: 2890]
- **Aufgabe:** Frequenzauflistung mit ▶ `Frequencies`
- **Zusatzaufgabe:** Sortieren Sie die Frequenzauflistung zuerst nach Häufigkeit, dann alphabetisch.

1.4 Einfache Kollokationen (einfache Frequenzen)

Mit welchen Nomen kommt das attributive Adjektiv *deutsch* vor? Testen Sie zunächst die einfachen Suchanfragen und sichten Sie die Ergebnisse. Ein einzelnes `+` steht in der CEQL-Syntx für ein beliebiges Wort (warum funktioniert `*` hier nicht?). Bei der dritten Anfrage verlangen Sie zusätzlich, dass das auf *deutsch* folgende Wort ein Substantiv ist. Wortarten (*parts of speech*, POS) können mit der Notation `_{POS}` gesucht werden (z.B. `_{N}` für Nomen). Die einfachen POS-Tags, die Sie hier verwenden können, sind auf der Online-Hilfeseite erläutert.

- **Anfrage 1:** `deutsche* *`; `sort order=unsorted` [R: 2890]
- **Anfrage 2:** `deutsche* +`; `sort order=unsorted`
- **Variante:** `deutsche* +`; `sort order=ascending`
- **Anfrage 3:** `deutsche* _{N}` (oder: `deutsche* +_{N}`) [R: 2640]
- **Aufgabe:** Frequenzen der Wortpaare sichten (▶ `Frequencies`)

1.5 Komposita

Welche Komposita mit dem Kopf *-gesetz* gibt es ? Die beiden Anfragen ergeben sehr ähnliche Resultate. Wo unterscheiden Sie sich?

- **Anfrage 1:** `*gesetz`; `mode=literal` [R: 595]
- **Anfrage 2:** `+gesetz`; `mode=case-folded`

Ein Nachteil bei diesem Ansatz ist, dass nur die Grundformen der Komposita gefunden werden, nicht aber flektierte Formen wie *Grundgesetzes* oder *EU-Gesetze*. (Warum funktioniert die Anfrage `*gesetz*` nicht?) Sie können eine Lemma-Suche durchführen, indem Sie das Suchwort in geschweifte Klammern `{...}` einschließen. Lassen Sie sich anschließend die Häufigkeiten anzeigen (`sort order=unsorted`). Wie werden die Komposita übersetzt? Finden Sie lexikalisierte Übersetzungen, halb lexikalisierte, Spontanbildungen?

- **Anfrage:** `{+gesetz}`; `mode=case-folded` [R: 1369]
- **Aufgabe 1:** Häufigkeiten anzeigen lassen (`sort order=unsorted`; ▶ `Frequencies`)
- **Aufgabe 2:** Übersetzungen sichten (z.B. *Grundgesetz, Antiterrorgesetz, Religionsgesetz*) — dazu in der Frequenzliste auf das jeweilige Wort klicken (in der Display-Option müssen die entsprechenden Sprachen aktiviert sein)
- **Aufgabe 3:** Alternative Sortierung (`sort order=reverse desc.`; ▶ `Frequencies`)
- **Aufgabe 4:** Übersetzungen sichten

1.6 Wortlisten ("Word list"-Modus)

Suchen Sie nun nach Gesetzes-Komposita im *Word List*-Modus (Link im Hauptmenü oben). Wie wird *Grundgesetz* übersetzt?

- **Auswahl:** `Word List`
- **Einstellungen:** `Word list for: German`, `attribute=lemma`, `sort=by frequency`, `literal`, `filter=+gesetz`; ▶ `Make Word List`
- **Aufgabe:** Sichten der Übersetzungen (Klick auf `T` neben Wort)

1.7 Disjunktion in einfachen Anfragen

Gehen Sie wieder zurück in den *Simple Query*-Modus. Sie wollen wissen, ob im Europarl-Korpus häufiger von *Atomkraft* oder von *Kernkraft* die Rede ist. Hierzu verwenden Sie eine Disjunktion in der Anfrage.

- **Anfrage 1:** `[Kern,Atom]kraft` [R: 329]
- **Anfrage 2:** einschließlich Komposita mit `[Kern,Atom]kraft+` [R: 1220]
- **Aufgabe:** Häufigkeiten anzeigen lassen (`sort order=unsorted`)

1.8 Syntax: Folge von Adjektiven

Welche Folgen mit zwei unmittelbar aufeinander folgenden Adjektiven gibt es im Korpus? Zunächst suchen Sie über Wortartenfilter (Adjektive: A). Sichten Sie das Ergebnis. **Kommentar:** Die automatische Wortartenannotation verwechselt oft Adjektive und Adverbien, da sie im Deutschen sehr ähnlich im Satz verteilt sind.

Um die Genauigkeit (*precision*) zu erhöhen verwenden wir unseren alten Trick, die Flexionsendung direkt anzusprechen. Diesmal sollten Sie über einen Operator jeden beliebigen Stamm erlauben und die Flexionsendungen einzeln auflisten. Kombinieren Sie dieses Muster mit dem POS-Tag für Adjektive.

Da die Trefferzahl immer noch sehr groß ist, suchen Sie anschließend nach Sequenzen mit drei Adjektiven. Um die Anfrage übersichtlicher zu gestalten verwenden Sie einen numerischen Wiederholungsoperator: `_{A} _{A} _{A}` kann auch als `( _{A} ){3}` geschrieben werden.

- **Anfrage 1:** `_{A} _{A}` [R: > 100 000]
- **Anfrage 2:** `*[e,er,es,em,en]_{A} *[e,er,es,em,en]_{A}` [R: > 100 000]

- **Anfrage 3:** `( *[e,er,es,em,en]_{A} ){3}` [R: 2376]
  Sichtung mit `sort=unsorted`
- **Anfrage 4:** `( *[e,er,es,em,en]_{A} ){4}` [R:29]
- **Zusatzaufgabe:** Falls Ihnen das STTS-Tagset bekannt ist: Formulieren Sie die Anfrage über ein geeignetes STTS-Tag, das im Gegensatz zu den bisherigen POS-Angaben nicht in `{...}` eingeschlossen wird. Müssen Sie auch jetzt noch die Adjektiv-Endungen angeben, um "saubere" Ergebnisse zu erhalten?

1.9 *Wie es im Buche steht* – Realisierung des Dativ *-e*

Zunächst verschaffen Sie sich einen ersten Eindruck und testen die spezielle Kombination von *in* + *Buch*.

- **1. Test:** `im Buche`; `mode=literal` [R: 2]
- **2. Gegentest:** `im Buch` [R: 6]

Da das Korpus keine morphologische (Kasus-)Information enthält, müssen Sie indirekt Dativ einfordern: z.B. über eindeutige Artikel und Präpositionen.

- **Anfrage 3:** `[einem,dem,am,beim,im,vom,zum] ( _{A} )* *e_{N}`; [R: 31 910]
  `mode=case-folded`; `sort=unsorted`

Die Ergebnisse sehen nicht schlecht aus. Aber die Präzision ist nicht gut: Treffer 1 und 4 z.B. sind *false positives*. **Kommentar:** Bessere Ergebnisse wären zu erwarten, wenn man zusätzlich verlangen könnte, dass das zugehörige Lemma nicht auf *-e* endet. Hierzu benötigen wir eine mächtigere Anfragesprache, die auch einen Negationsoperator enthält: die CQP-Syntax.

Bevor Sie in den *CQP Mode* wechseln, scrollen Sie an das Seitenende. Dort wird Ihnen die Übersetzung Ihrer aktuellen Anfrage in CQP-Syntax angezeigt.

- **Aufgabe:** Kopieren Sie die CQP-Variante Ihrer Suchanfrage.

2 Europarl – CQPweb ("CQP Mode")

Zunächst bleiben Sie im Europarl-Webinterface. Wechseln Sie nur in den CQP-Modus (durch Klick auf `CQP Mode` im Hauptmenü oben).

- **URL:** http://cogsci.uni-osnabrueck.de/~korpora/ws/CQPdemo/Europarl/
  ▶ `CQP mode`
- **Hilfe:** http://cogsci.uni-osnabrueck.de/~korpora/ws/CQPdemo/Europarl/help-cqp.html

2.1 Realisierung des Dativ *-e* – Anfragen in CQP-Syntax

Kopieren Sie Ihre letzte Anfrage aus Abschnitt 1.9 in das Suchfenster. Verstehen Sie die einzelnen Komponenten? Ergänzen Sie dann den negativen Filter und überprüfen Sie die Resultate. Was passiert, wenn man den "Dativfilter" weglässt?

- **Anfrage 1:** [R: 31 836]
  `[word="(einem|dem|am|beim|im|vom|zum)"%c] ([pos="ADJ."])* [word=".*e" & pos="N.*"]`
- **Anfrage 2:** [R: 18 197]
  `[word="(einem|dem|am|beim|im|vom|zum)"%c] ([pos="ADJ."])*`
  `[word=".*e" & lemma!=".*e" & pos="N.*"]`
  Einstellungen: `lang=German`; deaktivieren: `EN,FR,ES,IT,NL`
- **Aufgabe:** Häufigkeiten anschauen; ggf. `normalized` deaktivieren
- **1. Zusatzaufgabe:** Was passiert, wenn Sie den "Dativfilter" weglassen?
- **2. Zusatzaufgabe:** Ersetzen Sie den "Dativfilter" durch einen STTS-Wortartenfilter für Artikel und das Muster *-m* für die Wortendung [R: 18 185]
  `[pos="ART|APPRART" & word=".*m"] [pos="ADJ."]* [word=".*e" & lemma!=".*e" & pos="N."]`

- **Variante:** eine kompaktere Formulierung ist `[pos=".*ART" & word=".*m"] ...`

Klicken Sie auf `Distribution`, um eine Aufteilung der Treffer nach Jahr und Originalsprache zu sehen. Es fällt dabei auf, dass die Dativformen mit *-e* (fast) ausschließich in Texten vorkommen, die ins Deutsche übersetzt wurden. **Kommentar:** Der beobachtete Zeitabschnitt ist zu klein, um der Aufteilung nach Jahren aussagekräftige Tendenzen zu entnehmen. In Abschnitt 3 wechseln Sie auf die DWDS-Korpora, die eine Zeitspanne von 100 Jahren umfassen und eine bessere diachrone Darstellung erlauben.

2.2 *Gegenseitiges Vertrauen* – Wechsel zu CQPweb

Welche Adjektive treten mit dem Nomen *Vertrauen* auf? Wechseln Sie zu dem CQPweb-Interface in Osnabrück. Falls dabei eine Sicherheitswarnung erscheint, bestätigen Sie bitte die Ausnahme und akzeptieren das Sicherheitszertifikat des Servers.

- **URL:** https://cogsci.uni-osnabrueck.de/~korpora/ws/cqpweb/
- Zugangsinformationen (Login + Passwort) werden im Tutorium ausgeteilt

Wichtige Anmerkung: Wenn man auf `New Query` klickt, wird der Such-Modus auf *Simple Query* zurückgesetzt. Daher ist es oft bequemer, den "Zurück"-Knopf des Web-Browsers zu verwenden. (Bei *Simple Query* können Sie die gleiche CEQL-Syntax wie im Europarl-Webinterface verwenden. Wir werden hier jedoch ausschließlich mit CQP-Syntax arbeiten.)

Suchen Sie zunächst nach dem Nomen Vertrauen. Nutzen sie dann die Kollokations-Suche von CQPweb mit verschiedenen Filteroptionen. Zuerst untersuchen Sie den linken Kontext des Nomens *Vertrauen* (Fenster: 3 Token), in dem Sie relevante Adjektive erwarten. Anschließend untersuchen Sie den rechten Kontext von *Vertrauen*. Hier erwarten Sie typischerweise Nomen, die gehäuft mit *Vertrauen* verwendet werden.

- **Anfrage:** `[lemma="Vertrauen"]`
  Einstellung: `Query mode = CQP syntax` [R: 5451]
- **Schritt 1:** ▶ `Collocations`
  `Include: Lemma, POS;` ▶ `Create collocation database`
  `Collocation based on: Lemma;`
  `Collocation window from: 3 to the left;`
  `Collocation window to: 1 to the left;`
  ▶ `[Submit changed parameters] Go!`
- **Schritt 2:** zusätzlich Filter `and/or tag: ADJA` aktivieren; ▶ `Go!`
- **1. Resultate:** Klick auf eines der Adjektive – zeigt detaillierte Statistiken zu dieser Kollokation
- **2. Resultate:** Klick auf Zahl in Spalte *Observed collocate frequency* – zeigt alle Belege

Untersuchung des rechten Kontexts:

- **Schritt 3:** "Zurück"-Knopf (um zur Kollokationsliste zurückzukehren)
  `Collocation based on: Lemma;`
  `Collocation window from: 1 to the Right;`
  `Collocation window to: 3 to the Right;`
  ▶ `[Submit changed parameters] Go!`
- **Schritt 4:** zusätzlich Filter `and/or tag: NN` aktivieren
- **Aufgabe:** Details für *Verbraucher* anschauen. In welchem Abstand von *Vertrauen* tritt das Wort am häufigsten auf?
- **Zusatzaufgabe:** Suchen Sie nach Kollokationen mit dem Wort *deutsch* (Filter: Adjektive).

2.3 *Ohne Dirigent(en)* – Können "nackte" Substantive flektiert werden?

In diesem Abschnitt untersuchen wir die Gültigkeit der Regel, dass Substantive nicht flektiert werden, wenn sie ohne Artikel oder Adjektiv verwendet werden. Um unnötige *false positives* zu vermeiden, wählen wir den eindeutig erkennbaren Kontext einer Präpositionalphrase. Die erste Anfrage resultiert in viele Pluralnomen, daher ergänzen wir einen Filter, der typische Pluralendungen aussortiert.

- **Anfrage 1:** `[pos="APPR"] [pos="NN" & lemma != word]` [R: 233 783]
- **Anfrage 2:** `[pos="APPR"] [pos="NN" & lemma!=word & word!=".*(n|e|er)"]` [R: 16 879]

Die Ergebnisse der letzten Anfrage enthalten immer noch alle dem POS-Tagger unbekannten Nomen, die in Europarl als `<unknown>` lemmatisiert sind. Wir schließen diese nun explizit aus (das Flag `%c` ist wegen Inkonsistenzen in der Europarl-Annotierung erforderlich):

- **Anfrage 2:** [R: 9 390]
  `[pos="APPR"] [pos="NN" & lemma!=word & lemma!="<unknown>"%c & word!=".*(n|e|er)"]`
- **Aufgabe:** ▶ `Frequency breakdown`: Sichten Sie die ersten 80 Fälle

3 DWDS-Suche: ein diachrone Perspektive

Wechseln Sie auf die Seite des Digitalen Wörterbuchs der Deutschen Sprache.

- **URL:** http://beta.dwds.de/
- **Hilfe:** http://beta.dwds.de/help/

3.1 Vorbemerkungen

Die Anfragesprache des DWDS unterscheidet sich deutlich von CQP. Die wichtigsten Unterschiede sind im Folgenden aufgelistet. Wie die Korpora oben ist das DWDS aber mit STTS-Tags annotiert.

- Im DWDS wird per Default nach Lemmata gesucht; eine Wortform muss extra mit `@` markiert werden!
- Sobald man *Wildcards* (Platzhalter) nutzt, wird jedoch automatisch nach Wortformen gesucht. Z.B. müsste `*blau` eigentlich logischerweise `@*blau` heißen. D.h. also: Man kann nicht mit Wildcards nach Lemmata suchen (und unsere Flexionsanfragen werden komplizierter).
- Man kann in jedem Wort maximal eine Wildcard verwenden.
- Sucht man nach Wort-Sequenzen (d.h. nach mehreren Wörtern, die unmittelbar aufeinander folgen), muss die Anfrage in doppelte Anführungszeichen eingeschlossen werden.
- Disjunktionen werden immer "satzweise" ausgewertet. Der Suchausdruck `A || B` ("A oder B") bedeut also: A oder B kommen (irgendwo) im Satz vor. Eine weitere Folge ist, dass z.B. der Ausdruck `"A B (C || D)"` so nicht formuliert werden kann, sondern als `"A B C" || "A B D"` geschrieben werden.
- Nach POS-Annotationen sucht man mit dem Ausdruch `$p=<POS>`, gegebenenfalls gefolgt von einer Angabe des Lemmas: `$p=<POS> with <Lemma>` (oder auch: `<Lemma> with $p=<POS>`).

3.2 Einfache Wortsuche

Das DWDS bietet verschiedene Korpora (und Wörterbücher) für die Suche an. In der Voreinstellung ("Standardsicht") ist das DWDS-Kernkorpus und das ZEIT-Korpus ausgewählt. Testen Sie die Anfragen von oben auf diesen beiden Korpora (Trefferangaben für das Kernkorpus). Dabei die Anführungszeichen bei Anfragen nach mehreren Wörtern nicht vergessen!

- **Anfrage 1:** `deutsch` [R: 140.345]
- **Anfrage 2:** `@deutsch` [R: 3.673]
- **Anfrage 3:** `"@auf @gut @deutsch"` [R: 10]

3.3 Attributives Adjektiv und Kollokationen mit Substantiven

Wie suchen Sie nach Belegen für *deutsch* als attributives Adjektiv? Mit welchen Substantiven tritt *deutsch* im Kernkorpus auf? Dazu fügen wir wieder POS-Information zur Anfrage hinzu. Als eine weitere Variante können Sie angeben, dass zwischen zwei Wörtern n andere Wörter stehen. In Anfrage 3 sind es 0 Wörter, d.h. die Anfrage ist zu Anfrage 2 äquivalent.

- **Anfrage 1:** `deutsch with $p=ADJA` [R: ?]
- **Anfrage 2:** `"deutsch with $p=ADJA $p=NN"` [R: 89.575]
- **Anfrage 3:** `"deutsch with $p=ADJA #0 $p=NN"` [R: 89.575]

Geben Sie nun wieder die Anfrage ein, die nach dem Lemma *deutsch* sucht. Im Fenster "DWDS-Wortprofil" sehen Sie in Form einer Wortwolke die wichtigsten Kollokationen des Lemmas *deutsch*. Durch Klicken auf eine der Funktionen, die als "Relationenfilter" oberhalb der Wortwolke gelistet sind, können Sie die Kollokationen einschränken. Klicken Sie auf "Attribut", um ausschließlich Kollokationen zu erhalten, in denen *deutsch* in einer Attribut-Relation vorkommt. Testen Sie auch die Funktion "Beiordnung". Warum gibt es für die anderen Funktionen kein Ergebnis?

Noch eine zweite Kollokationssuche: *Gegenseitiges Vertrauen*. Überlegen Sie zuerst, welche Kollokationen mit *Vertrauen* Ihnen sonst noch einfallen, z.B. Verben, die mit Vertrauen in der Funktion als Subjekt, als Objekt eine Kollokation bilden? Geben Sie dann ein:

- **Anfrage:** `Vertrauen` [R: 48.844]
- **Aufgabe:** Schauen Sie sich dazu auch die Wortwolke an und testen Sie verschiedene Relationenfilter.

3.3.1 Komposita

Kommen wir nun zu den Komposita. Suchen Sie jetzt nur noch auf dem Kernkorpus, da es balanciert ist, nach Zeit (Jahre 1900–2000) und nach Genre (Zeitung, Belletristik, Wissenschaft, Gebrauchsliteratur, (etwas) gesprochene Sprache).

Das DWDS erlaubt es auch (wie CQP), die Frequenz der Treffer im zeitlichen Verlauf darzustellen. Dazu am linken Rand auf "+ Panel" klicken, "Statistik hinzufügen" auswählen und dann die "Kernkorpus-Wortverlaufsstatistik" hinzufügen. Das neue Panel wird unten angefügt.

- **Anfrage 1:** `*gesetz` [R: 24.295]
- **Aufgabe 1:** Testen Sie verschiedenen Sortierungen unter `Darstellungsoptionen` (nach Datum, Dokumentlänge, ...).
- **Aufgabe 2:** Schauen Sie sich die `Wortverlaufsstatistik` an, die Ihnen die Trefferzahl pro Genre und Jahrzehnt anzeigt.

3.4 Syntax: Sequenz von Adjektiven

Wir wiederholen die Suche von oben nach mindestens 4 aufeinanderfolgenden attributiven Adjektiven:

- **Anfrage:** `"$p=ADJA $p=ADJA $p=ADJA $p=ADJA"` [R: 138]

3.5 Realisierung des Dativ *-e* – im DWDS

Da man im DWDS keine Platzhalter auf Lemmata anwenden kann und auch keine Disjunktion auf Token-Ebene möglich ist, wird die Anfrage etwas komplexer als in CQP. Der Ausdruck `#1` erlaubt ein optionales Wort zwischen dem Artikel und dem Substantiv.

- **Anfrage:** `"@einem #1 @Buche" || "@dem #1 @Buche" || "@am #1 @Buche" || "@beim #1 @Buche" || "@im #1 @Buche" || "@vom #1 @Buche" || "@zum #1 @Buche"` [R: ?]
- **Aufgabe:** Da es sich um eine alte Form des Dativs handelt, ist hier die diachrone Entwicklung besonders interessant: die Wortverlaufsstatistik anschauen.
- **Aufgabe:** Suchen Sie sich ein anderes Nomen und vergleichen Sie die Ergebnisse miteinander.

4 Chunks und morphologische Merkmale

Seit einigen Jahren ist eine wachsende Anzahl effizienter automatischer Annotierungswerkzeuge verfügbar, deren Möglichkeiten über die Wortartenannotierung und Lemmatisierung herkömmlicher Tagger hinausgehen. Insbesondere ist es inzwischen möglich geworden, auch große Korpora syntaktisch zu analysieren und mit morphologischen Merkmalen zu annotieren. In der vorliegenden Übung wollen wir uns ansehen, wie sich solche Annotationen in Korpusabfragen nutzen lassen.

Als Datenbasis dienen uns deutsche Zeitungstexte aus den 1990er-Jahren, die mit partiellen syntaktischen Analysen (sog. *Chunking*) und morphologischen Merkmalen annotiert sind. Beim Chunking handelt es sich um eine Erkennung lokaler Phrasenstruktur. Im Gegensatz zu den heute verbreiteten statistischen Parsern werden dabei keine unzuverlässigen Attachment-Entscheidungen vorgenommen. Auch bei der morphologischen Analyse wurde auf eine statistische Desambiguierung verzichtet. Stattdessen sind alle Hypothesen der Morphologiekomponente in einer mengenwertigen Annotationsebene gespeichert. Es findet lediglich eine partielle Desambiguierung durch Kongruenz innerhalb der erkannten Chunks statt, deren Ergebnis folgerichtig auf Chunk-Ebene annotiert wurde.

Auf die Zeitungskorpora kan über eine Variante des aus Übung 1 bekannten Europarl-Webinterfaces zugegriffen werden. Die erweiterten Annotationen sind allerdings nur in CQP-Anfragen verwendbar, nicht in der einfachen CEQL-Syntax. Geben Sie zunächst in Ihrem Browser die untenstehende Adresse des Webinterfaces ein.

- **URL:** https://cogsci.uni-osnabrueck.de/~korpora/ws/CQPdemo/HGC/frames-cqp.html
- Sie können hier dieselben Anmeldedaten (Login + Password) wie in Übung 2.2 verwenden.

Machen Sie sich zunächst mit dem Webinterface vertraut, indem Sie einfache Suchanfragen mit Einzelwörtern und regulären Ausdrücken ausführen. Vergleichen Sie dabei auch die vier verschiedenen Tageszeitungen, die abgefragt weren können (die unten aufgeführten Trefferzahlen beziehen sich stets auf die *Frankfurter Rundschau*).

- **Einstellungen:** Mit der neuen Option `phrases` können Sie die vom Chunker gefundenen Nominal- und Präpositionalphrasen (im folgenden als NP und PP bezeichnet) anzeigen lassen.
- **Aufgabe:** Probieren Sie die verschiedenen Anzeigemodi aus (▶ in manchen Modi wird zusätzliche Information angezeigt, wenn der Mauszeiger auf einer öffnenden Klammer ruht).

Sie werden feststellen, dass die angezeigten Chunks rekursiv verschachtelt sind. Im Gegensatz zur englischen Sprache, für die der Chunk-Begriff zunächst eingeführt wurde, ist im Deutschen keine sinnvolle nicht-rekursive Definition von Chunks möglich. Wir werden im folgenden nur mit *maximalen* Chunks, die in keinen größeren Chunk desselben Typs eingebettet sind, arbeiten. Diese werden im Webinterface in Fettdruck dargestellt.

- Detaillierte Informationen zu der erweiterten Anfragesyntax für Chunks und morphologische Merkmale finden Sie in Abschnitten 4 und 5 des *CQP Query Language Tutorial*.
  ▶ http://cogsci.uni-osnabrueck.de/~korpora/ws/CWBdoc/CQP_Tutorial/

4.1 Zugriff auf Chunks

Im Zuge der automatischen Annotierung werden die erkannten Chunks in Form von XML-Tags wie `<np>` (Beginn eines Nominalchunks) und `</np>` (Ende eines Nominalchunks) in den Text eingefügt. Solche Tags können auch in CQP-Anfragen eingesetzt werden und finden jeweils eine passende Chunk-Grenze. Werden Tags des gleichen Chunk-Typs paarweise verwendet (wie in Anfrage 3 unten), so muss es sich um Anfang und Ende desselben Chunks handeln. Diese XML-Notation kann auch für Satzgrenzen (`<s>` ... `</s>`) verwendet werden.

- **Anfrage 1:** `<np> [pos = "APPR"]` [R: 21 644]
- **Aufgabe:** Gibt es tatsächlich Nominalphrasen, die mit einer Präposition beginnen?
- **Anfrage 2:** `[pos = "APPR"] </s>` [R: 704]
- **Anfrage 3:** `<np> [pos != "ADJ."]{10,} </np>` [R: 2942]
  ▶ lange Nominalphrasen ohne Adjektive
- **Aufgabe für Wetterfrösche:** Können Sie eine CQP-Anfrage formulieren, die NPen mit vier oder mehr Adjektiven findet? Welche besondere Schwierigkeit stellt sich hierbei?

`<s> ... </s>`	Satz
`<np> ... </np>`	Nominalphrase (NP)
`<pp> ... </pp>`	Präpositionalphrase (PP)
`<ap> ... </ap>`	Adjektivphrase (AP) mit Argumenten und Adjunkten
`<vc> ... </vc>`	Verbalkomplex
`<cl> ... </cl>`	Nebensatz (systematisch erfasst sind v. a. Relativsätze)
`<mw> ... </mw>`	Mehrwortlexeme (*mehr als*, *bis zu*, *ein paar*, ...)

Tabelle 1: In den Zeitungskorpora annotierte XML-Tags.

Tabelle 1 stellt die in den Zeitungskorpora verfügbaren XML-Tags zusammen.

Die meisten Chunks sind mit zusätzlichen Informationen über die jeweilige Phrase annotiert, z. B. ihrem lexikalischen Kopf (lemmatisiert). Auf diese Annotationen kann mit Hilfe einer speziellen Syntax für Start-Tags zugegriffen werden. Beispielsweise findet

```
<np_h = "Linguistik">
```

den Anfang einer NP mit dem Kopf *Linguistik*. Dabei können analog zu Tokenmustern auch reguläre Ausdrücke sowie die Modifikatoren `%c` und `%d` verwendet werden. Um die kompletten Chunks zu finden, muss das passende End-Tag `</np_h>` mit angegeben werden, dazwischen stehen beliebig viele Token.

- **Anfrage 1:** `<np> []* [lemma = "Linguistik"] []* </np>` [R: 8]
- **Aufgabe:** Anfrage 1 versucht, NPen mit dem Kopf *Linguistik* ohne Verwendung der expliziten Annotation zu finden. Prüfen sie, ob *Linguistik* tatsächlich bei allen Treffern Kopf der NP ist. Falls nicht: Wo könnte das Problem liegen?
- **Anfrage 2:** `<np_h = "Linguistik"> []+ </np_h>` (*Achtung:* hier nicht `</np>`) [R: 6]
- **Aufgabe:** Führen Sie Anfrage 1 und Anfrage 2 für einige weitere Kopflemmata aus. Welche der beiden Anfragen wird schneller bearbeitet?
- **Anfrage 3:** `<np_h = ".*gesetz"> []+ </np_h>` [R: 4781]
- **Quizfrage:** Können Sie erklären, warum die Anfrage `<np_h = "Linguistik">` keine Treffer findet?

Sehr nützlich sind auch die Köpfe von Verbalkomplexen, wie die folgende Anfrage verdeutlicht.

- **Anfrage:** `<vc_h = ".+setzen"> []* [lemma = "setzen"] []* </vc_h>` [R: 3374]

NPen und andere Chunks sind darüber hinaus mit bestimmten Merkmalen wie „Eigenname“, „in Anführungszeichen“, „Maßangabe“, „Zeit/Datum“, usw. ausgezeichnet. Diese können mit dem Tag `<np_f>` (und analog `<pp_f>`, `<ap_f>`, ...) abgefragt werden. Eine Besonderheit ist hierbei, dass es sich um mengenwertige Annotationen handelt, da ein Chunk mehrere Merkmale besitzen kann. Solche Merkmalsmengen werden an Stelle von `=` mit den speziellen Vergleichsoperatoren `contains` (die Menge *enthält* das gesuchte Merkmal) und `matches` (alle Merkmale müssen mit dem angegebenen regulären Ausdruck *übereinstimmen*) abgefragt.

- **Anfrage:** `<np_f contains "street"> []+ </np_f>` [R: 66 850]

Eine Aufzählung der wichtigsten Merkmale verschiedener Chunk-Typen findet sich im Anhang des *CQP Query Language Tutorial*.

4.2 Morphologische Annotation

Die Zeitungstexte sind auf Tokenebene mit morphologischen Merkmalen wie Kasus, Genus und Numerus annotiert, die für nominale Kongruenz relevant sind. Da im Deutschen häufig Synkretismen vorliegen, werden alle möglichen Merkmalskombinationen in Form einer Menge angegeben. Die Elemente der Menge folgen dabei stets dem Schema

Kasus : *Genus* : *Numerus* : *Artikeltyp*

Kasus	`Nom`, `Gen`, `Dat`, `Akk`
Genus	`M`, `F`, `N`
Numerus	`Sg`, `Pl`
Artikeltyp	`Def`, `Ind`, `Nil`

Beispielsweise sind die Wortformen *der* (als Artikel) und *Stoffe* folgendermaßen annotiert:

```
der     |Dat:F:Sg:Def|Gen:F:Pl:Def|Gen:F:Sg:Def
        |Gen:M:Pl:Def|Gen:N:Pl:Def|Nom:M:Sg:Def|
Stoffe  |Akk:M:Pl:Def|Dat:M:Sg:Def|Gen:M:Pl:Def|Nom:M:Pl:Def
        |Akk:M:Pl:Ind|Dat:M:Sg:Ind|Gen:M:Pl:Ind|Nom:M:Pl:Ind
        |Akk:M:Pl:Nil|Dat:M:Sg:Nil|Gen:M:Pl:Nil|Nom:M:Pl:Nil|
```

Mit regulären Ausdrücken lassen sich nun Muster angeben, die ein oder mehrere Merkmale herausgreifen. Werden diese mit dem Operator `contains` kombiniert, so handelt es sich um mögliche Eigenschaften der gefundenen Wörter; mit dem Operator `matches` müssen diese Eigenschaften eindeutig sein.

- **Anfrage 1:** `[pos = "ART" & agr contains "Gen:.:Sg:.*"]` [R: > 100 000]
- **Anfrage 2:** `[pos = "ART" & agr matches "Gen:.:Sg:.*"]` [R: > 100 000]
- **Aufgabe:** Vergleichen Sie die Artikel, die von den beiden Anfragen gefunden werden. Sie können dazu mit der Schaltfläche `Frequencies` eine Häufigkeitsliste erstellen.

Die morphologischen Merkmale lassen sich teilweise durch Kongruenz innerhalb von NPen desambiguieren. So bleibt im obigen Beispiel für die NP *der Stoffe* nur eine konsistente Merkmalskombination `Gen:M:Pl:Def` übrig. Diese partiell desambiguierte morphologische Information ist auf Chunk-Ebene annotiert und kann mit den Methoden aus Abschnitt 4.1 abgefragt werden.

- **Anfrage:** `<np_agr matches "Gen:.:Pl:.*"> []+ </np_agr>` [R: 43 345]
- **Aufgabe:** Finden Sie mit dieser Anfrage auch die NP *der Stoffe*? Suchen Sie gezielt danach, um nicht alle 43 345 Treffer durchsehen zu müssen. Nutzen Sie die Anzeigeoptionen um zu erklären, warum Sie mehr Belege für *der Stoffe* finden, wenn Sie `np_agr` jeweils durch `np_agr1` ersetzen.

4.3 Eine Fingerübung *wider besseren Wissens*

Laut Duden steht die Präposition *wider* stets mit Akkusativ, z. B. *wider den tierischen Ernst* und *wider besseres Wissen*. Ziel dieser Übung ist, Belege für Verwendungen von *wider* mit anderen Kasus zu finden. Nutzen Sie hierzu die bisher gelernten Funktionen. Möglicherweise sind auch die negierten Vergleichsoperatoren `not matches` und `not contains` hilfreich.

- **Anfrage:** `"wider"%c <np_agr matches "Dat:.*"> []+ </np_agr>` [R: 3]
- **Aufgabe:** Wandeln Sie die Anfrage ab, um mehr relevante Beispiele zu finden.
- **Aufgabe:** Sie werden unter den Ergebnissen u. a. auch *das Für und Wider eines Arguments* entdecken. Wie können Sie die Anfrage verändern, um das Nomen *Wider* von den Treffern auszuschließen?

4.4 *Weil man sagt das halt so* – Verbzweitstellung nach *weil*

Weil ist eine unterordnende Konjunktion, die eigentlich einen Verbendsatz einleitet:

> *Sie ist nicht nach Hause gefahren, weil sie Kopfschmerzen hatte.*

Umgangssprachlich wird weil aber auch mit Verbzweitsätzen verwendet:

> *Sie ist nicht nach Hause gefahren, weil sie hatte Kopfschmerzen.*

- **Aufgabe:** Finden sich in den Zeitungskorpora auch Belege für *weil* mit Verbzweitstellung? Nutzen Sie die Chunk-Annotationen, um die Satzstellung nach *weil* möglichst explizit und vollständig zu modellieren.
  ▶ `<cl>`-Chunks helfen Ihnen dabei, von Relativsätzen modifizierte NPen zu erkennen.

5 Das STTS-Tagset

Tag	Beschreibung	Beispiele
ADJA	attributives Adjektiv	*[das] große [Haus]*
ADJD	adverbiales oder prädikatives Adjektiv	*[er fährt] schnell, [er ist] schnell*
ADV	Adverb	*schon, bald, doch*
APPR	Präposition; Zirkumposition links	*in [der Stadt], ohne [mich]*
APPRART	Präposition mit Artikel	*im [Haus], zur [Sache]*
APPO	Postposition	*[ihm] zufolge, [der Sache] wegen*
APZR	Zirkumposition rechts	*[von jetzt] an*
ART	bestimmter oder unbestimmter Artikel	*der, die, das, ein, eine, ...*
CARD	Kardinalzahl (Ordinalzahlen sind als ADJA getaggt)	*zwei [Männer], [im Jahre] 1994*
FM	Fremdsprachliches Material	*[Er hat das mit "] A big fish [" übersetzt]*
ITJ	Interjektion	*mhm, ach, tja*
KOUI	unterordnende Konjunktion mit *zu* und Infinitiv	*um [zu leben], anstatt [zu fragen]*
KOUS	unterordnende Konjunktion mit Satz	*weil, daß, damit, wenn, ob*
KON	nebenordnende Konjunktion	*und, oder, aber*
KOKOM	Vergleichskonjunktion	*als, wie*
NN	normales Nomen	*Tisch, Herr, [das] Reisen*
NE	Eigennamen	*Hans, Hamburg, HSV*
PDS	substituierendes Demonstrativpronomen	*dieser, jener*
PDAT	attribuierendes Demonstrativpronomen	*jener [Mensch]*
PIS	substituierendes Indefinitpronomen	*keiner, viele, man, niemand*
PIAT	attribuierendes Indefinitpronomen ohne Determiner	*kein [Mensch], irgendein [Glas]*
PIDAT	attribuierendes Indefinitpronomen mit Determiner	*[ein] wenig [Wasser], [die] beiden [Brüder]*
PPER	irreflexives Personalpronomen	*ich, er, ihm, mich, dir*
PPOSS	substituierendes Possessivpronomen	*meins, deiner*
PPOSAT	attribuierendes Possessivpronomen	*mein [Buch], deine [Mutter]*
PRELS	substituierendes Relativpronomen	*[der Hund ,] der*
PRELAT	attribuierendes Relativpronomen	*[der Mann ,] dessen [Hund]*
PRF	reflexives Personalpronomen	*sich, einander, dich, mir*
PWS	substituierendes Interrogativpronomen	*wer, was*
PWAT	attribuierendes Interrogativpronomen	*welche [Farbe], wessen [Hut]*
PWAV	adverbiales Interrogativ- oder Relativpronomen	*warum, wo, wann, worüber, wobei*
PAV	Pronominaladverb	*dafür, dabei, deswegen, trotzdem*
PTKZU	*zu* vor Infinitiv	*zu [gehen]*
PTKNEG	Negationspartikel	*nicht*
PTKVZ	abgetrennter Verbzusatz	*[er kommt] an, [er fährt] rad*
PTKANT	Antwortpartikel	*ja, nein, danke, bitte*
PTKA	Partikel bei Adjektiv oder Adverb	*am [schönsten], zu [schnell]*
TRUNC	Kompositions-Erstglied	*An- [und Abreise]*
VVFIN	finites Verb, voll	*[du] gehst, [wir] kommen [an]*
VVIMP	Imperativ, voll	*komm [!]*
VVINF	Infinitiv, voll	*gehen, ankommen*
VVIZU	Infinitiv mit *zu*, voll	*anzukommen, loszulassen*
VVPP	Partizip Perfekt, voll	*gegangen, angekommen*
VAFIN	finites Verb, aux	*[du] bist, [wir] werden*
VAIMP	Imperativ, aux	*sei [ruhig !]*
VAINF	Infinitiv, aux	*werden, sein*
VAPP	Partizip Perfekt, aux	*gewesen*
VMFIN	finites Verb, modal	*dürfen*
VMINF	Infinitiv, modal	*wollen*
VMPP	Partizip Perfekt, modal	*gekonnt, [er hat gehen] können*
XY	Nichtwort, Sonderzeichen enthaltend	*3:7, H2O, D2XW3*
$,	Komma	*,*
$.	Satzbeendende Interpunktion	*. ? ! ; :*
$(	sonstige Satzzeichen; satzintern	*– [ ] ( )*